Supporting Information

Zhang *et al.* 10.1073/pnas.0812600106

SI Materials and Methods

Clone Library Construction for Sanger Sequencing. We prepared PCR mixes in 25-μL reaction volumes composed of 12.5 μL of 2.5× PCR 5′ Mastermix (Eppendorf) with 2 mM $MgCl_2$, 0.2 mM dNTPs, and 0.1 μg of each primer. We performed PCR in an Eppendorf Mastercycler EP using an initial denaturing step at 95 °C for 2 min, followed by 25 cycles of 94 °C for 45 s, 54 °C for 45 s, and 72 °C for 1.5 min, with a final elongation step at 72 °C for 7 min. To minimize PCR artifacts associated with differential template abundance (1), we pooled individual PCR reactions starting with 0.1, 1, and 2 μL of DNA. We excised PCR amplicons of correct size from the agarose gels using a clean razor blade and recovered them with Montage gel extraction spin columns (Millipore). We cloned the purified PCR products using the plasmid vector pCR4-TOPO TA (Invitrogen) for sequencing according to the manufacturer's instructions. We grew the transformed TOP10 *Escherichia coli* cells overnight on LB agar plates containing 100 μg/mL ampicillin, and then extracted plasmids through a semiautomated alkaline lysis method described previously (2) and sequenced the inserts using an ABI 3730xl capillary sequencer with the same 8f primer used in PCR. After removing vector sequences and low-quality ends with the SeqMan Pro software (DNASTAR), we used nearest-alignment space termination (NAST) to align the trimmed Sanger sequencing reads. We then imported the NAST-aligned sequences into an ARB database (greengenes.arb, version 23, May 2007) using the "FASTA with gaps" filter. After manually correcting the alignment in the editor ARB_EDIT4 (3) based on known 16S rRNA secondary structures, we added the sequences into a backbone tree ("tree_all" with 137,916 near–full-length reference sequences) with the "LanemaskPH" filter, which excludes highly variable regions that may overestimate the tree's branch lengths. We kept only sequences obtained in this study in the final phylogenetic tree.

16S rDNA V6 Pyrosequencing. To pool and sort multiple samples in a single 454 GS-FLX run, we designed unique tags of 5 nucleotides to identify each sample. The resulting forward primer was a fusion of the 454 life science adaptor A, the tag, and 967f (5′-gcctccctcgcgccatcag-"tag"-CAACGCGAAGAACCTTACC-3′). The reverse primer was unchanged. The PCR conditions were 94 °C for 2 min, 25 cycles of denaturation at 94 °C for 30 s, 57 °C annealing for 45 s, and 72 °C for a 1-min extension, followed by a final extension at 72 °C for 2 min. We purified the PCR products using QiaQuick spin columns (Qiagen) to remove excess primer dimers and dNTPs and measured the concentrations of PCR amplicons with a NanoDrop spectrophotometer. After a sequencing run and base calling, we sorted the sequences by unique tags using the 454 script "sfffile" to separate and group all data, and then trimmed the sequences using the 454 script "sffinfo" for downstream analysis.

Automated Data Analysis Pipelines Using Pyrosequencing Data. We developed 2 automated analysis pipelines. The first pipeline was aimed at assigning 454 sequences to phylotypes and included 4 steps. First, 454 reads were preprocessed to remove ambiguous and short sequences, all sequences having mismatches with the PCR primers, and all sequences having fewer than 50 nucleotides after the proximal primer (unless they reached the distal primer). These filtering steps eliminated all of the sequences with more than 1 ambiguity (N). Second, each remaining sequence was compared through similarity searches, using the program BLASTN (4), against a reference database including 44,011 nonidentical V6 sequences extracted from 119,480 bacterial rRNA genes. Third, the BLAST output was parsed to pull out up to 150 best hits having alignments longer than 57 bp, which were aligned using the program "MUSCLE" (5) set to the following parameters: -diags and -maxiters 2. Finally, the program "QuickDist," modified from QuickTree (6, 7), was used to parse the alignments to pull out sequence(s) from the reference database having the minimum distance to the original query. Automation of the workflow was facilitated by a set of Perl scripts that initiate each of the appropriate programs, parse the results, and properly summarize the output data. After the aforementioned 4-step matching pyrosequencing tag queries to the reference sequences, we used the Ribosomal Database Project's Classifier 2.0 (8) to assign taxonomy.

The second pipeline estimated species richness with ecological statistical tools. We performed multiple sequence alignment using the MAFFT program and the PartTree algorithm, with the parameters -parttree and -retree1. The MAFFT alignment was converted to the Pfam Stockholm format, which served as an input to QuickDist for generating a distance matrix.

Real-Time QPCR. To quantify *Bacteria*, *Archaea,* and the archaeal subgroups, we performed 16S rRNA gene-targeted quantitative real-time PCR (QPCR) with either SYBR-green (for *Bacteria*) or TaqMan detection (for *Archaea, Methanobacteriales*, *Methanomicrobiales,* and *Methanosaetaceae*). The *Archaea*-specific primers and probes were Arc787f (5′-ATTAGATACCCSBGTAGTCC-3′), Arc1059r (5′-GCCATGCACCWCCTCT-3′), and Arc915probe (5′-AGGAATTGGCGGGGGAGCAC-3′). The *Methanobacteriales*-specific primers and probe were MBT857f (5′-CGWAGGGAAGCTGTT AAGT-3′), MBT1196R (5′-TACCGTCGTCCACTCCTT-3′), and MBT929-probe (5′-AGCAC CACAACGCGTGGA-3′). The *Methanomicrobiales*-specific primers and probe were MMB282F (5′-ATCGRTACGG GTTGTGGG-3′), MMB832R (5′-CACCTAACGCRCATHGTTTAC-3′), and MMB749-probe (5′-TYCGACAGTGAGGRACGAAAGCTG-3′). The *Methanosaetaceae*-specific primers and probe were MST702F (5′-TAATCCTYGARGGACCACCA-3′), MST862R (5′-CCTAC GGCACCRACMAC-3′), and MST753-probe (5′-ACGGCAAGGGACGAAAGCTAGG-3′).

We prepared plasmid DNA standards from representative 16S rRNA gene clones of target organisms. For *Bacteria,* we used an uncultured *Xanthomonas* sp. (*Gammaproteobacteria*), and for *Archaea* and *Methanobacteriales,* we used *Methanobacterium formicicum*. For *Methanomicrobiales* and *Methanosaetaceae,* we used plasmid standards DQ301905 and AY 570685, respectively. A 20-μL PCR contained 10 μL of the iQ Supermix (BioRad), 1 μL of each primer (0.5 μM final concentration), 0.04 μL of FAM-labeled probe (TaqMan, Applied Biosystems), and 2 μL of template DNA. QPCR was performed in an Eppendorf Realplex gradient cycler with a procedure that combines annealing and extension. Specifically, the initial denaturing was at 94 °C for 3 min, followed by 45 cycles of 94 °C for 15 s and 60 °C for 15 s. For every primer pair set, we ran triplicate QPCR reactions for each of the 9 individuals in the nw, ob, and gb groups. We calculated the copy number of the 16S rRNA gene per gram of wet stool as follows: # per gram wet stool = $q/2 \times D \times (200\ \mu L)/(0.2\ g)$, where q is the detected copy numbers from 2 μL of diluted template, D is the dilution factor, 200 μL is the elution

volume in genomic DNA extraction, and 0.2 g is the wet weight of stool used for DNA extraction.

DGGE. We generated PCR amplicons for DGGE analysis with a published nested PCR protocol (9). In brief, we used *Archaea*-specific primers 20f (10) and 1492r (11) in the first round of PCR and PARCH340f and PARCH519r (9) in the second round of PCR. After cleaning the PCR products using QiaQuick spin columns, and we loaded (12 μL) and separated them using a D-CODE system (Bio-Rad) in 8% polyacrylamide gels prepared with 30% and 70% denaturant concentrations. Electrophoresis was performed at 60 °C under a constant voltage of 70 V for 16 h. The gels were stained with SYBR Green for 20 min, destained in water for 10 min, and then photographed.

1. Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* 64:3724–3730.
2. Kim H, *et al.* (2007) Comparative physical mapping between Oryza sativa (AA genome type) and O-punctata (BB genome type). *Genetics* 176:379–390.
3. Ludwig W, *et al.* (2004) ARB: A software environment for sequence data. *Nucleic Acids Res* 32:1363–1371.
4. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
5. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
6. Sogin ML, *et al.* (2006) Microbial diversity in the deep sea and the underexplored ''rare biosphere.'' *Proc Natl Acad Sci USA* 103:12115–12120.
7. Howe K, Bateman A, Durbin R (2002) QuickTree: Building huge neighbour-joining trees of protein sequences. *Bioinformatics* 18:1546–1547.
8. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
9. Ovreas L, Forney L, Daae FL, Torsvik V (1997) Distribution of bacterioplankton in meromictic Lake Saelenvannet, as determined by denaturing gradient gel electrophoresis of PCR-amplified gene fragments coding for 16S rRNA. *Appl Environ Microbiol* 63:3367–3373.
10. DeLong EF (1992) Archaea in coastal marine environments. *Proc Natl Acad Sci USA* 89:5685–5689.
11. Lane DJ (1991) in *Modern Microbiological Methods*, eds Stackebrandt E, Goodfellow M (Wiley, Chichester, UK), pp 115–175.

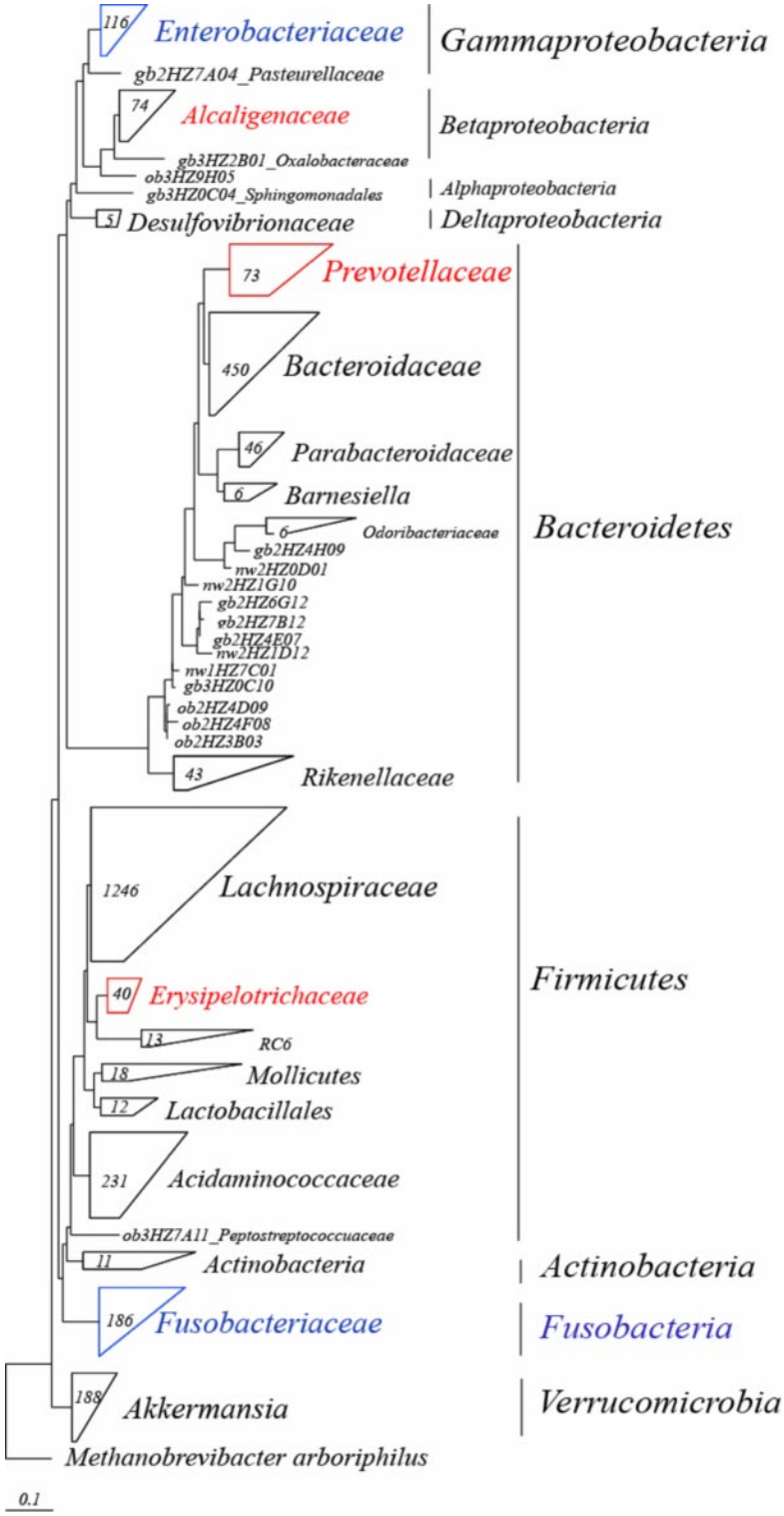

Fig. S1. Bacterial phylogenetic tree based on Sanger sequences. The number of sequences in a clade is indicated on the wedge. Sequences enriched in ob and gb individuals are indicated in red and blue, respectively.

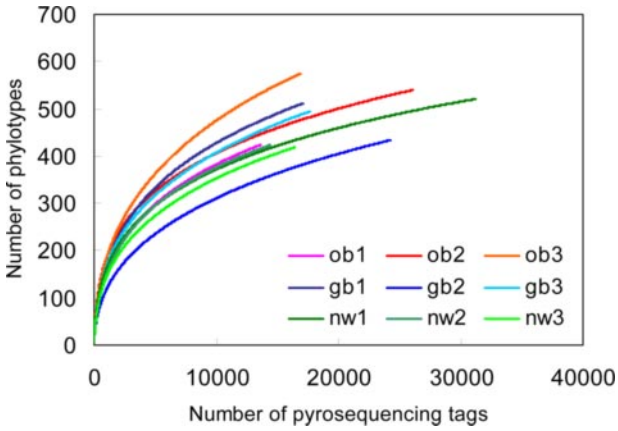

Fig. S2. Rarefaction curves calculated with the program Analytical Rarefaction 1.3, based on best matches in the V6RefDB.

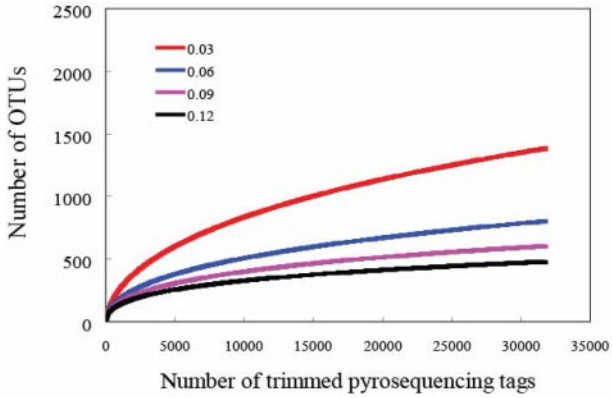

Fig. S3. DOTUR species-accumulation rarefaction analyses of pyrosequencing tags from sample nw1. Here 0.03 refers to 3% distance and corresponds to ''species'' level.

Table S1. Distance-based operational taxonomic units (OTUs) and species richness estimates

Sample ID	Trimmed tags	The observed species (OTUs at 0.03 difference)	Chao1 estimator of richness at 0.03 difference (95% CI)	ACE estimator of richness at 0.03 difference (95% CI)
nw1	31835	1,384	2206(2041,2412)	2217(2138,2305)
nw2	16260	873	1359(1239,1517)	1332(1277,1395)
nw3	16940	922	1531(1385,1724)	1505(1429,1592)
ob1	13963	777	1206(1097,1351)	1272(1206,1349)
ob2	26915	1,228	1915(1735,2146)	1924(1825,2036)
ob3	17334	1,107	1759(1617,1941)	1813(1734,1903)
gb1	17544	996	1678(1521,1881)	1652(1570,1745)
gb2	24794	891	1675(1525,1860)	1710(1648,1802)
gb3	18509	1,079	1723(1578,1910)	1708(1640,1784)

Unique tags, OTUs, and richness estimators were calculated using the DOTUR program, as described in *Materials and Methods*. OTU grouping at the 0.03 level refers to 3% distance and corresponds to bacterial species level. CI, confidence interval.